DOCUMENT RESUME

ED 118 590 TM 005 074

AUTHOR Donlon, Thomas F.
TITLE Establishing Appropriate Time Limits for Tests.
PUB DATE [Nov 73]
NOTE 22p.; Paper presented at the Annual Meeting of the Northeast Educational Research Associaiton (Ellenville, New York, November 1973)

EDRS PRICE MF-$0.83 HC-$1.67 Plus Postage
DESCRIPTORS *Statistical Analysis; *Test Construction; *Timed Tests

ABSTRACT

The implications of various time limits for tests of a fixed length or number of items seem obvious. If the time permitted is much too short, scores may bunch up at the low end of the potential score range, with a loss of potential variance and a general diminishing of the utility of the variance which is observed. If the time permitted is an intermediate value, the scores tend to become some mixture of power and speed. This paper proposes a simple technique for estimating the mean and standard deviation of the distribution of finishing times for a population of test takers. Given such values, a number of decisions concerning test specifications can be made. (Author/DEP)

Establishing Appropriate Time Limits for Tests

Thomas F. Donlon
Educational Testing Service
Princeton, N. J.

Presented at the Fall, 1973 Meeting
Northeast Educational Research Association

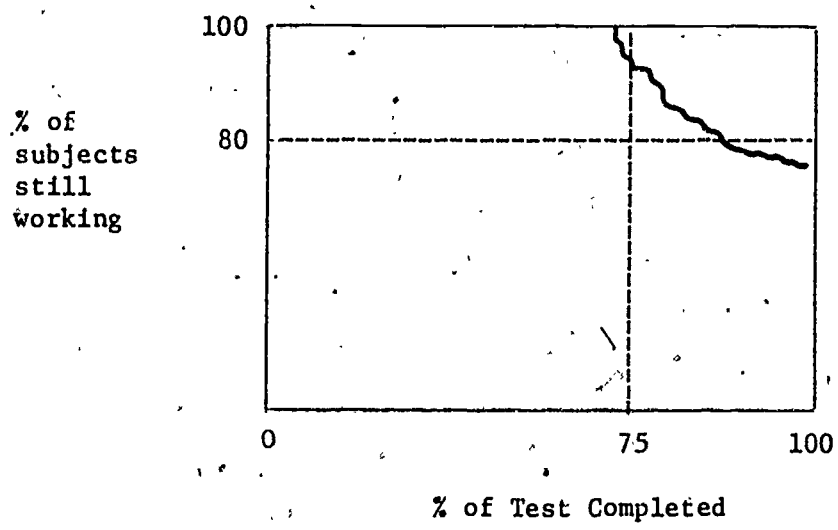
Establishing Appropriate Time Limits for Tests

The implications of various time limits for tests of a fixed "length" or number of items seems obvious. If the time permitted is much too short, scores may bunch up at the low end of the potential score range, with a loss of potential variance and a general diminishing of the utility of the variance which is observed. If the time permitted is an intermediate value, the scores tend to be some mixture of speed and power in the sense of Gulliksen (1950). Only with the most generous time limits do tests fully become power tests. Seldom are these generous times a practical possibility, however. Ordinarily time limits must meet the needs of the test administrator as much as the test taker. Under these conditions, some people finish, but some do not.

Such individual differences in the rate of work on objective tests complicate the meaning of the measures themselves, in the sense that they introduce factorial complexity. Accordingly, there is a general effort to make tests predominantly speed or power, and in most cases of measures of educational attainment the effort is made to create a power test. That is, the time limits are set so that most subjects complete the work.

No single technique for characterizing test speededness is widely established. Educational Testing Service has long focussed on three characteristics of the completion activity of the population taking the test: (a) the percent completing the test, (b) the percent completing 75% of the test, and (c) the test item at which approximately 80% of the total group are still working. These data are combined as criterion which, according to Swineford (1956) make a test speeded if (1) fewer than 100% of the candidates reach 75% of the items, and (2) fewer than 80% of the candidates finish 100% of the items.

Evans and Reilly (1972) used Swineford's criteria for speededness but introduced a graphic technique which plots the percent of candidates still working at various points in the test. Their own presentation used a base line of number of items. A more general approach would simply plot "percent of subjects still working" as a function of "percent of test worked on". An example would be



Such diagrams will characteristically exhibit the picture of a square with a "chunk" missing in the upper right hand corner. This is so because most tests do not show "dropout," in the sense of an appreciable percentage no longer reaching certain items, until the subjects are well into the test. Analogously, most tests are completed by most of the subjects. While the specific patterns of the function will vary, the curve will slope downward from the top to the right hand vertical axis. By Swineford's criteria, if this descending line lies within the region bounded by 75% of the test and 80% of the subjects, it is unspeeded. Departure from this region is a signal of speed.

Stafford (1971) has proposed a "speededness quotient," defined as

$$SQ = \frac{\Sigma U}{\Sigma W + \Sigma O + \Sigma U}$$

where U = the number of items Not Reached
O = the number of items Omitted
W = the number of items Wrong

According to Stafford, this index has some advantages over earlier indices proposed by Cronbach and Warrington (1951) and by Gulliksen (1950). These earlier suggestions formed ratios of variances of attempted items with total test scores on wrong answers. Stafford asserts that they were inordinately difficult to compute and to interpret.

All of these indices differ from the ETS approach in appraising variance in the number of items completed by considering variance in the number of errors. This is an important distinction between the two approaches. The results yielded will differ somewhat, too. Tests unspeeded by Swineford's criterion could show moderate speed by Stafford's criterion, approaching a value of about .25 as a limit. The general concern about speed and the widespread valuing of power, however, has ended to be psychometric, obscuring some other more practical considerations. Time limits tend to be set as they are because there are some real world needs. Time limits reflect these needs; so that a test conforms to a classroom period or to a three hc work session more from a need to conform to the institutional context than any psychometric factors.

The establishing of time limits introduces some practical problems, however, because of range of individual differences in work rates which is exhibited. Directors of large national programs have reported that in some tests or section

of tests, where large numbers of candidates work together in a room, some finish early, creating administrative difficulty for proctors. While the obvious solution to this problem is to give the candidates more to do, it is often unclear how much more material should be added, or by how much the time should be reduced. It is desirable then, to know something about the distribution of finishing times (or work rates) for various kinds of tests and varying composites of time allowance and test length.

This paper proposes a simple technique for estimating the mean and standard deviation of the distribution of finishing times for a population of test takers. Given such values, a number of decisions concerning test specifications can be made.

A frequent component of an item analysis is a description of the "drop-out" or failure to complete the test. This "dropout" is the percentage failing to reach the last item, and it can also be computed for other items. The graph adapted from Evans and Reilly, is based upon the dropout for a test.

For any subject who fails to complete the test we may estimate the time which would be required to finish. Thus, if a person completes one-half of the test in the prescribed amount of time, one can assume that the entire test would be completed in twice that time. If two-thirds is completed, then half again as much time is required. The basic assumption is that a person has a basic and consistent work rate, in the sense of items-per-minute.

The practical consequences of this assumption are that for anyone failing to finish the test we may estimate the time which would have been needed to finish. For those who complete the test, of course, we cannot develop an individual estimate. Although the number of "units" of work they accomplished is known, the time they worked is not. The only conclusion we may make is that the individual finished on or before the moment time was called.

However, if we can assume that the "time needed" (or, conversely the work rate) is normally distributed in the group being tested, we can associate a z-score with each "time-needed" score, by considering the proportion of the sample who exhibit scores equal to or greater than the one under consideration.

Consider a 30-item test with a 30-minute time limit. By the logic above, the following relationships exist

Items Reached	Not Reached	Items/ Minute	Time Needed
30	0	$\geq$1.00	$\leq$30.00 mins.
29	1	0.97	30.93 "
28	2	0.93	32.26 "
27	3	0.90	33.33 "
26	4	0.87	34.48 "
25	5	0.83	36.14
⋮	⋮	⋮	⋮
20	10	0.67	45.00 "

Suppose that the test is completed by 67% of the group. This indicates a z-score of 0.45 for this value. 67% of the group have work rates, expressed as items per minute, which equal or exceed a rate of 1.00. Suppose further that those reaching 29 or more of the items constitute 81% of the group. This indicates a z-score of 0.86, a time-needed score of 30.93.

Under the assumption of normality the information must be consistent for each of our two points. That is, a raw score (a "time needed" score stated in minutes) is a linear function of a z-score, a person's position in the normal distribution, and since the two points are supposedly linearly related, they should obey the function

$$S = (Z)\ (\sigma) + M$$

where S is the observed "time needed" score, Z is the normal curve z-score linked to the percentage with a score equal to or less than S, and σ and M are unknown group parameters.

Continuing the foregoing discussion, the assumptions yield two simultaneous equations for the data given above

$$30.00 = .45\sigma + M$$
$$30.77 = .86\sigma + M$$

they are solved to give values for M and σ of about 29.15 and 1.88 respectively. That is, if the assumptions are correct about a fairly steady work rate and a normal distribution of these rates, then the knowledge of two points on the distribution of Not Reached scores enables us to estimate the mean and standard deviation of the times needed.

With the knowledge that the average person completes the test in 29.15 minutes, and with the information that the standard deviation is equal to 1.88, one can estimate values of $M \pm 2\sigma$. These would be 32.91 minutes and 25.39 minutes. These data would suggest that the test is reasonably well timed. At the most, the very fastest workers are finishing about five minutes early, the very slowest might require some three minutes longer.

These hypothetical data consider only two points. In fact, an item analysis may yield a number of such points. Figure 1 shows the relationship between z-scores, as determined by the proportion of the population reaching a certain level, and the number of items reached. The essential linearity would seem to confirm the hypothesis of normality. As indicated, these data are for a 30-item test, administered in 30 minutes. The item material was "data sufficiency," a mathematical item type. The "items completable" entries show that on the average

approximately 32 items would be reached in 30 minutes, and that the standard deviation of item material finished would be about 7.8.

These data can be used to find the time needed for 30 items. The mean is 28 minutes, the standard deviation about 5.4 minutes. By this estimate about 16% of the population are finishing five minutes or more before time is called, and the very fastest workers, two standard deviations from the mean, are finishing 10 minutes early. Such a sizable group of early finishers might well constitute an administrative problem.

Figure 2 shows the plot for another 30-item, 30-minute pretest of mathematics material. Again, linearity seems observed. The lines in all figures in this paper are simply drawn in by hand, but there are few departures from a set of points which is easily connected.

Figure 3 shows data for a verbal pretest of 55 items in 30 minutes. Figure 4 presents a similar chart, but with a marked departure from linearity. Apparently the last five items in this pretest possessed some unusual characteristic which demanded more time. Instead of the approximately 75 items which might have been anticipated as an average completion only 50% or so of the group finished the test.

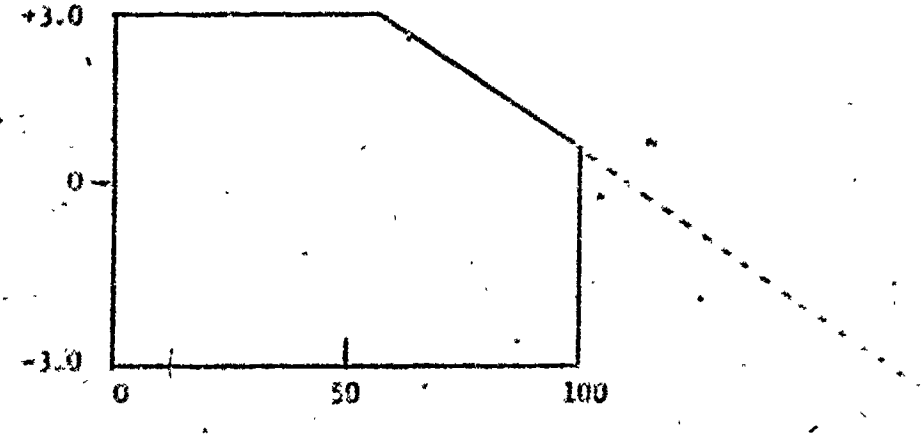
This is an unusual result. Figures 5-10 show the plots for six sections of two English Composition Tests. In each case the utility of the linear estimater can be defended, although points for values of z-above 2.0 tend to flatten off. This "flattening" may be due to the characteristics of the material or to the very small numbers of subjects.

Whether such data have practical value for predicting finishing times requires an empirical test. In a sense, they are surprising in that they summarize the experience of quite heterogeneous groups of people, in terms of ability

with heterogeneous groups of items in terms of difficulty but they would seem to conform rather closely to a model which anticipates a normally distributed items/minute work rate.

The "aberrant" plot in Figure 4 suggests the complexity of the "test speededness" concepts. It would appear that the last five items were found unusually time-consuming by the subjects, to a degree which ran counter to the average of such items. Is the test speeded? In some ways the introduction of difficult, time consuming later material may assist in the creations of orderly examination rooms. It may also affect the distribution of scores. It would be interesting to consider skewness data as they relate to speededness data of this type.

If the basic diagram of Reilly and Jackson is reconstituted so that the ordinate, or percent reaching axis is redefined as a z-score axis, the plots of such data should look approximately as follows:



While this diagram could be further modified to redefine the base line more clearly as a rate variable, the essential focus is established.

The linearity of the regression of "Z-completing" on "X-completed" invites extension, as indicated by the dotted line. Theoretically the slope of the line is determined by the collection of material which is incorporated into the test. A test which has a flat slope is more likely to be speeded, in some senses, than another. The task of the test constructor is to develop tests which have steep slopes. Swineford's criteria can now be restated in the new formulations: a test is unspeeded if the equation of its "dropout" line, as here defined, is

$$Z = .0864\,(100 - X) + .86$$

X = the percent of the test completed

or a line with steeper slope. The extension of this work to considerations posed by Stafford is clearly suggested.

References

Cronbach, L. J., & Warrington, W. G. Time limit tests: Estimating their reliability and degree of speeding. Psychometrika, 1951, 16, 167-188.

Evans, F. R., & Reilly, R. R. A study of speededness as a source of test bias. Journal of Educational Measurement, 1972, 9(2), 123-131.

Gulliksen, H. Theory of mental tests. New York: John Wiley and Sons, 1950.

Stafford, R. E. The speededness quotient: A new descriptive statistic for tests. Journal of Educational Measurement, 1971, 8(4), 275-278.

Swineford, F. Technical manual for users of test analyses. Statistical Report 56-42. Princeton, N.J.: Educational Testing Service, 1956.

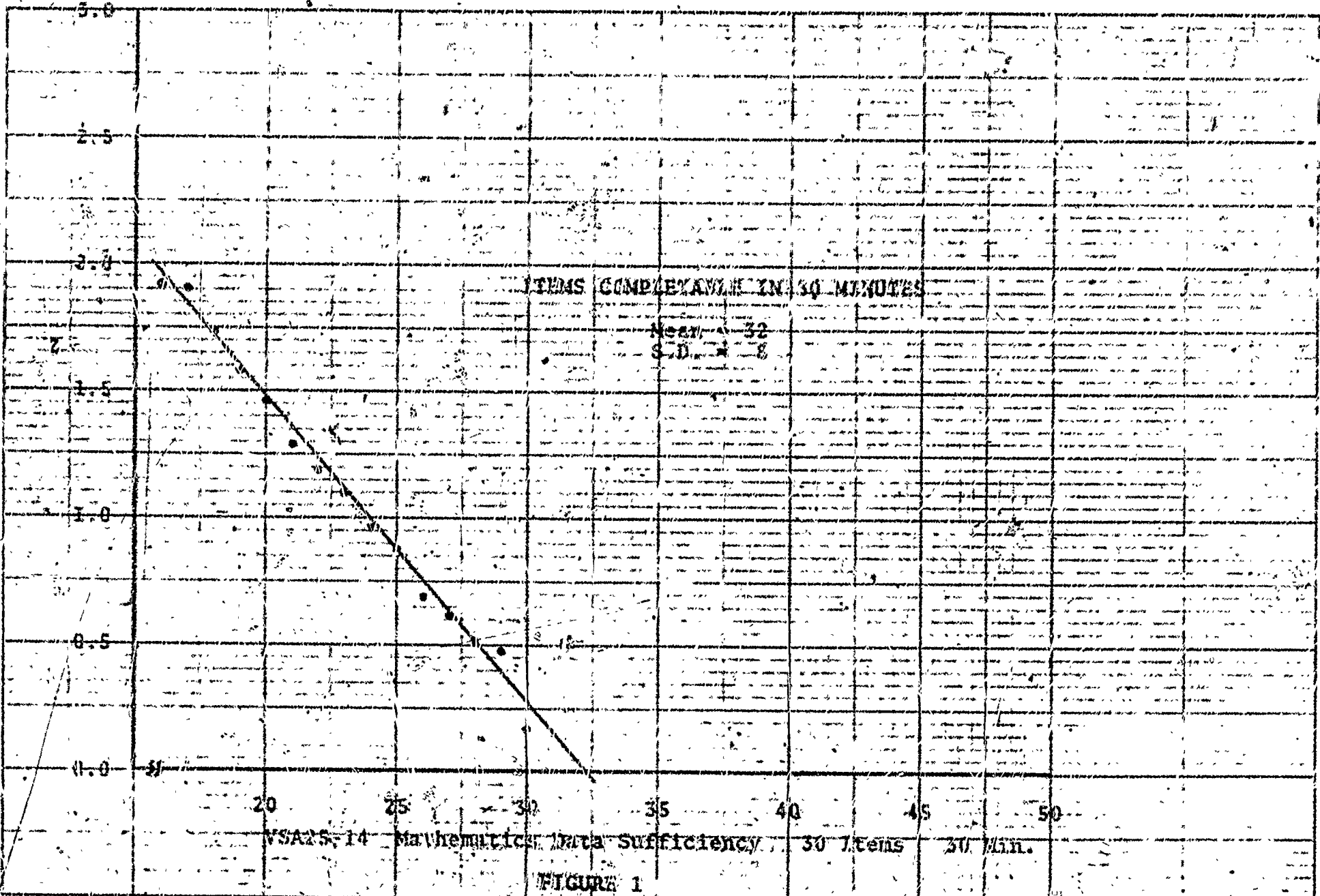
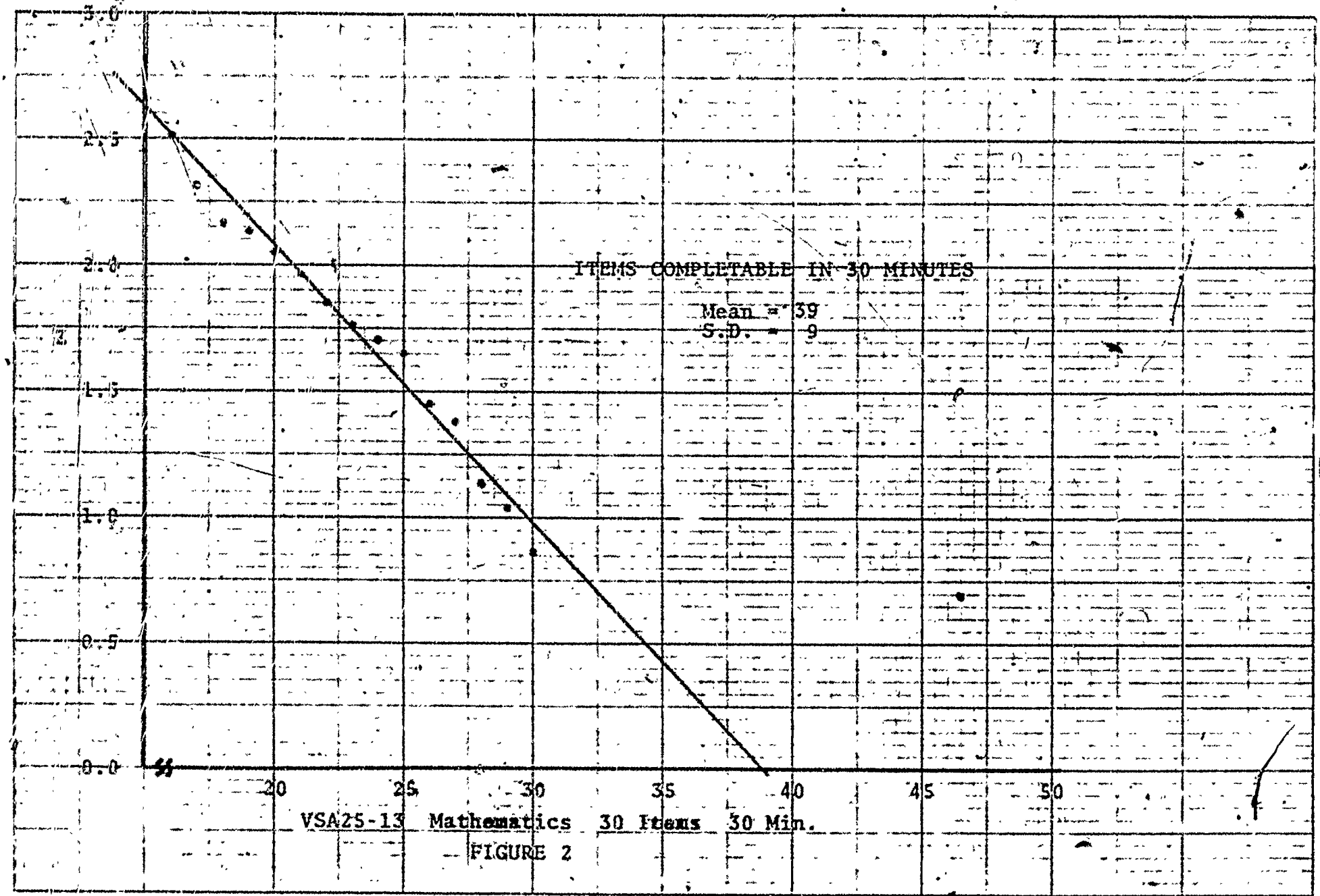
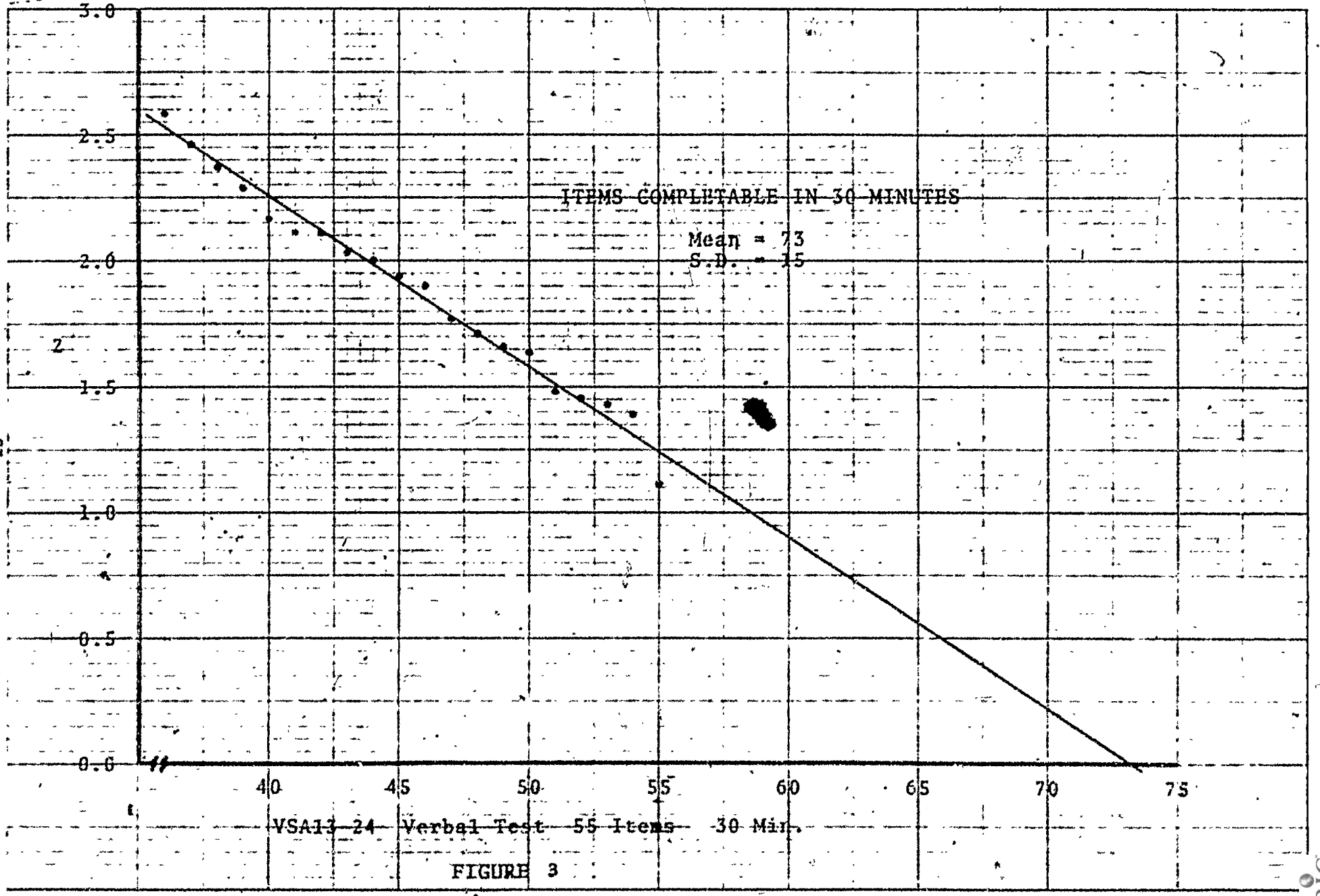

3.0
2.5
2.0
1.5
1.0
0.5
0.0
ITEMS COMPLETABLE IN 30 MINUTES
Mean = 32
S.D. = 8
20
25
30
35
40
45
50
VSAAS-14 Mathematics Data Sufficiency 30 Items 30 Min.

FIGURE 1



VSA25-13 Mathematics 30 Items 30 Min.

FIGURE 2



VSA13-24 Verbal Test 55 Items 30 Min.

FIGURE 3

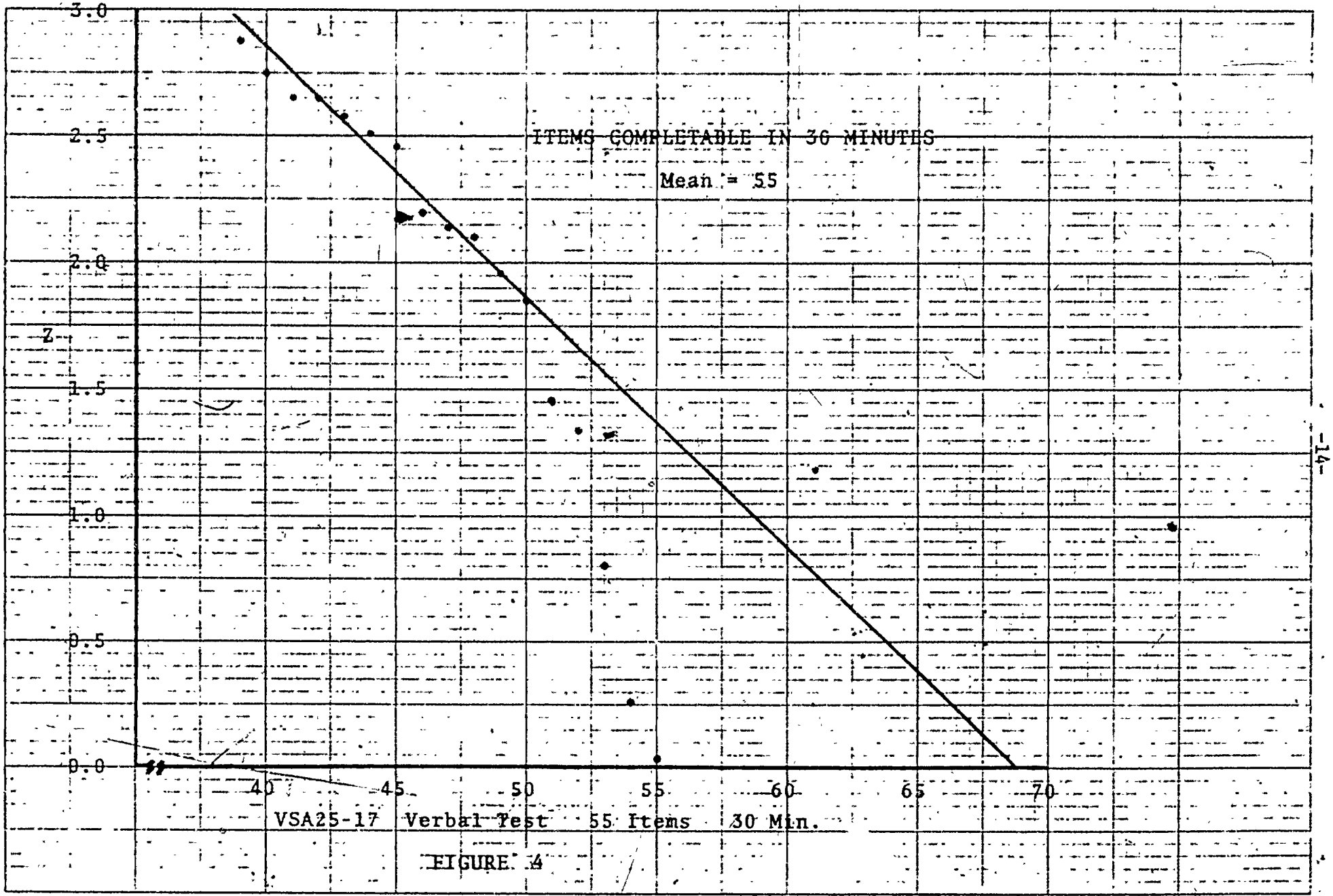
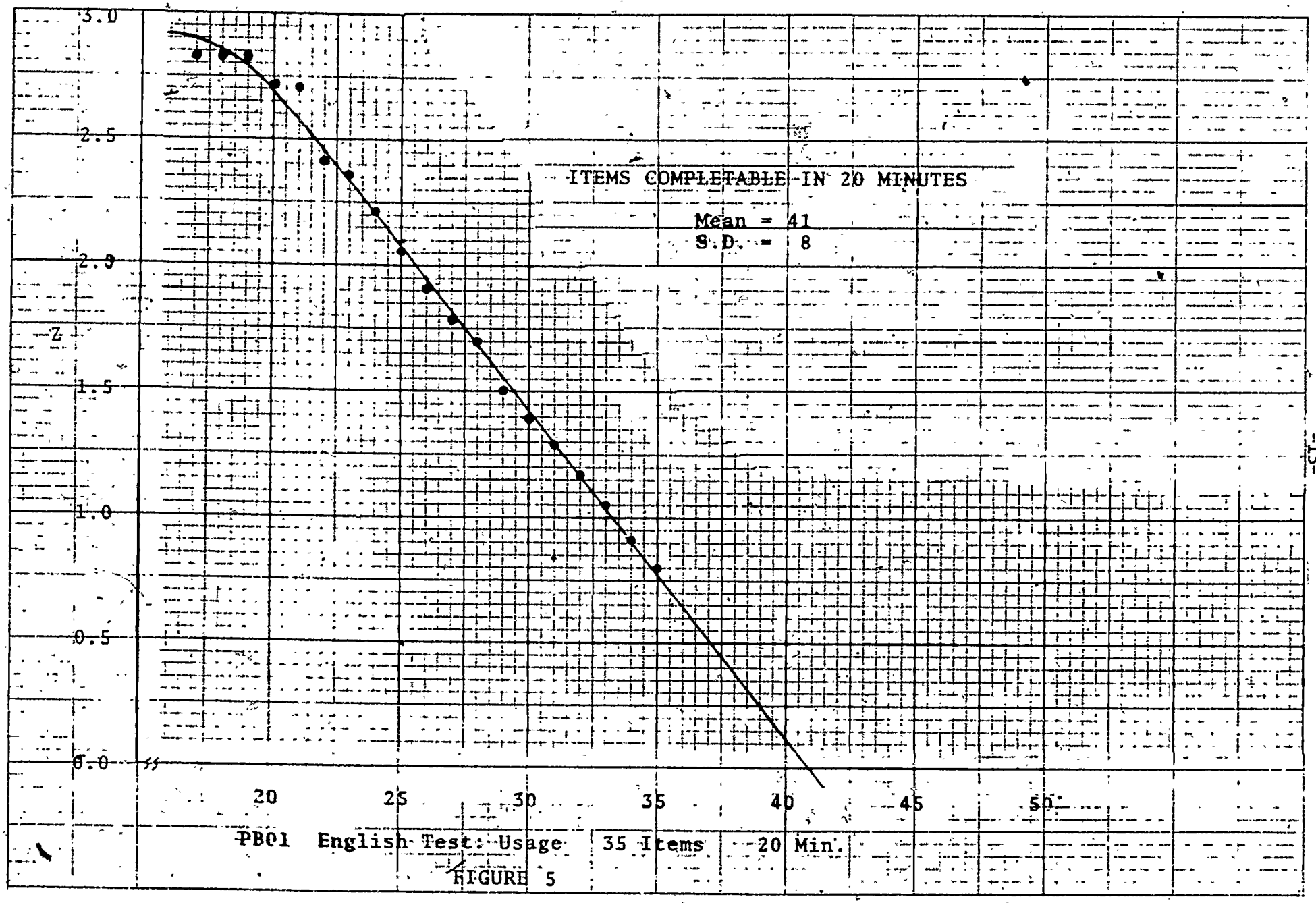
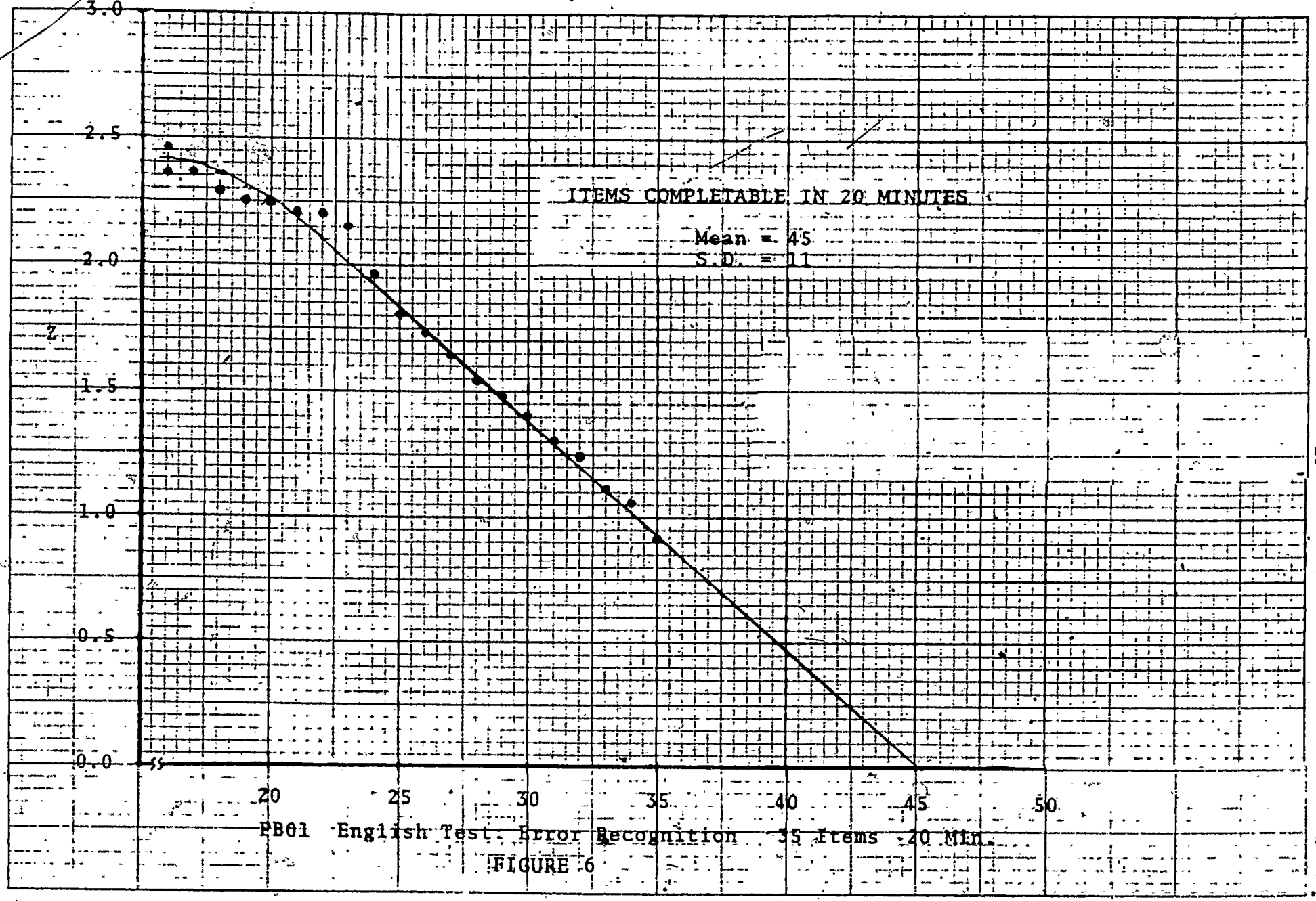

ITEMS COMPLETABLE IN 36 MINUTES
Mean = 55
Z
3.0
2.5
2.0
1.5
1.0
0.5
0.0
40
45
50
55
60
65
70
VSA25-17 Verbal Test 55 Items 30 Min.

FIGURE 4



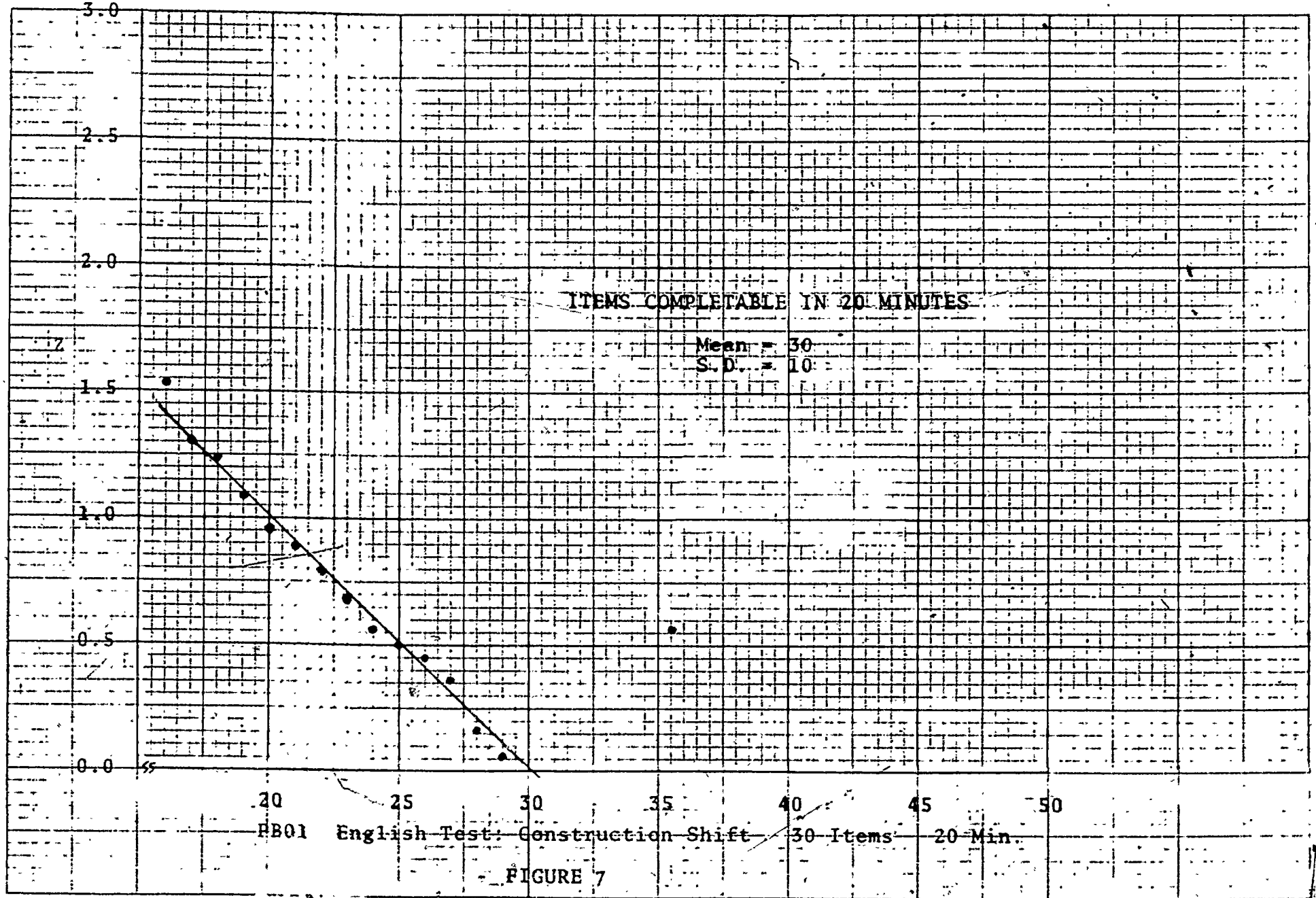
PB01 English Test: Usage 35 Items 20 Min.

FIGURE 5



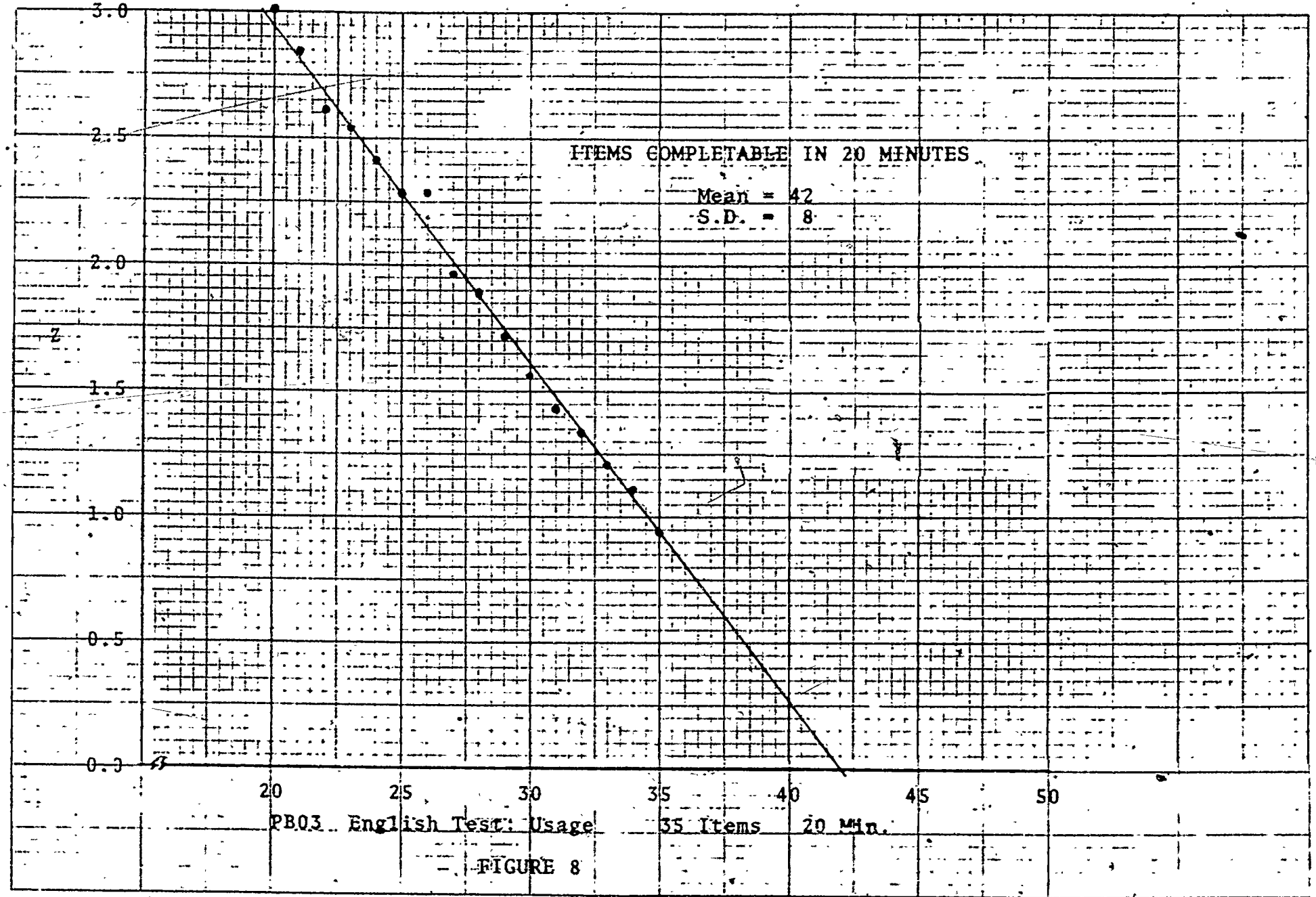
PB01 English Test. Error Recognition 35 Items 20 Min.

FIGURE 6



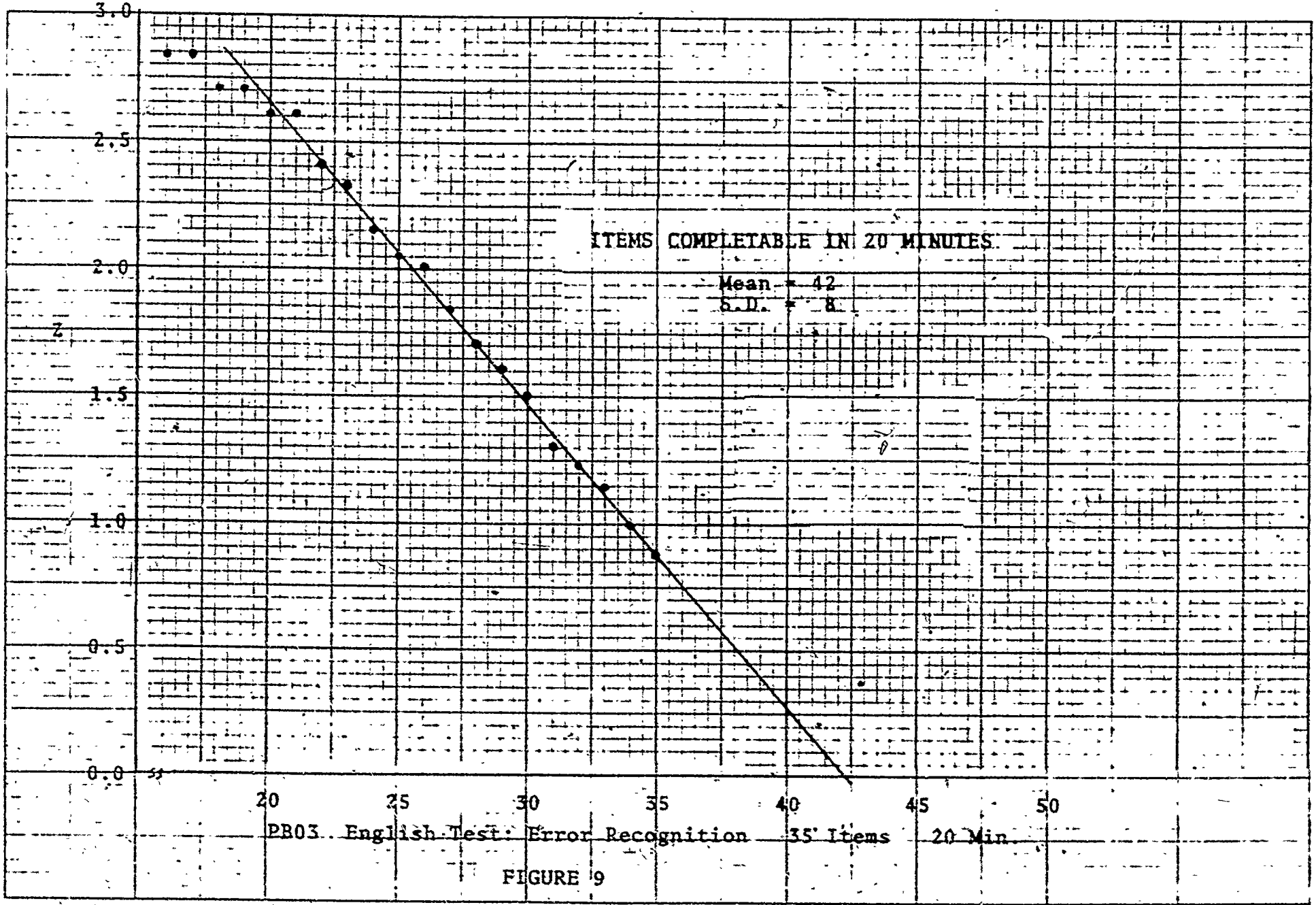
PB01 English Test: Construction Shift - 30 Items - 20 Min.

FIGURE 7



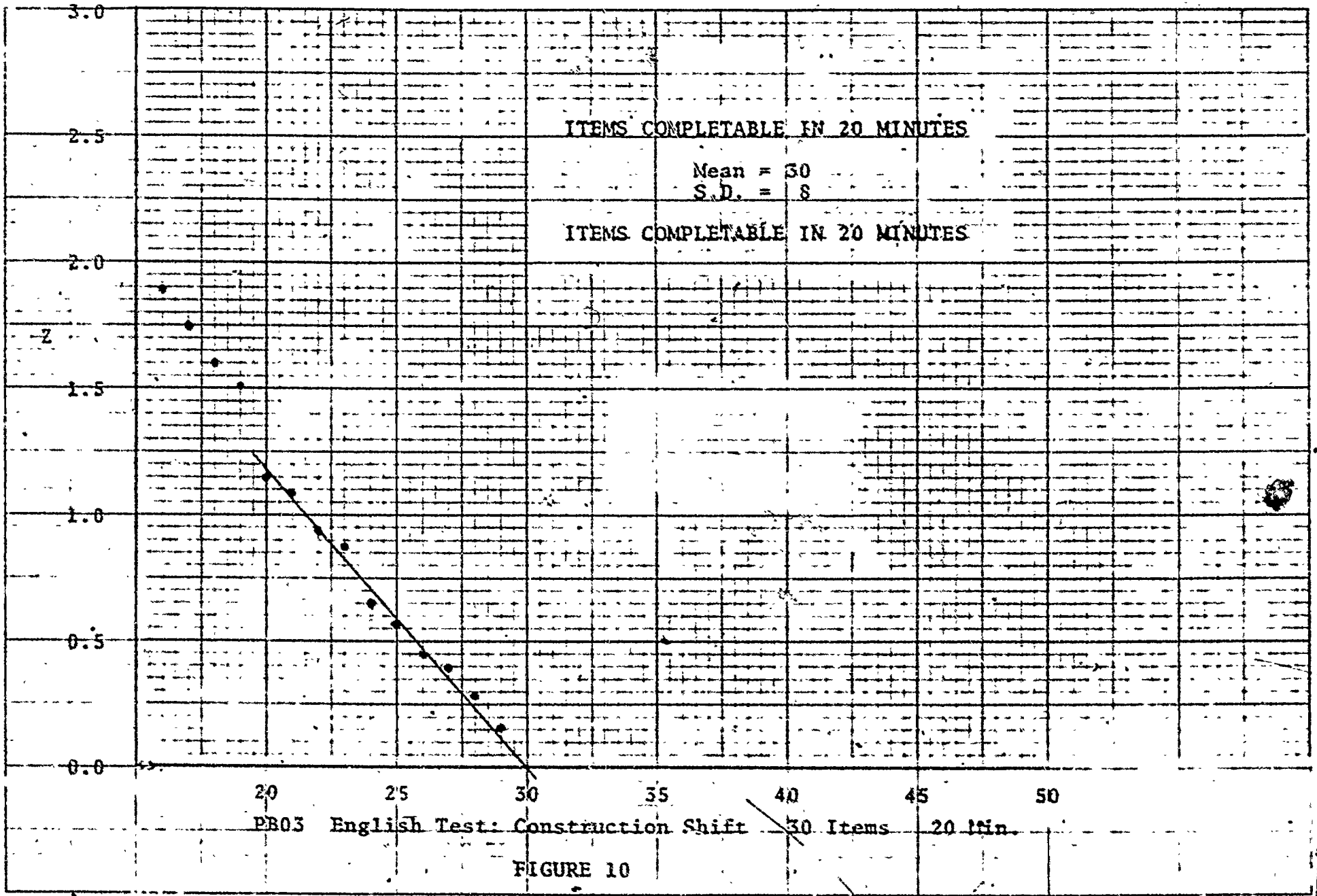
PB03 English Test: Usage 35 Items 20 Min.

FIGURE 8



PB03. English Test: Error Recognition 35 Items 20 Min.

FIGURE 9



PR03 English Test: Construction Shift 30 Items 20 Min.

FIGURE 10